Patent Application of

Lawrence Page

for

Method for Node Ranking in a Linked Database

CROSS-REFERENCES TO RELATED APPLICATIONS

This application claims priority from U.S. provisional patent application number 60/035,205 filed 01/10/97, which is incorporated herein by reference.

STATEMENT REGARDING GOVERNMENT SUPPORT

This invention was supported in part by the National Science Foundation grant number IRI-9411306-4. The Government has certain rights in the invention.

FIELD OF THE INVENTION

This invention relates generally to techniques for analyzing linked databases. More particularly, it relates to methods for assigning ranks to nodes in a linked database, such as any database of documents containing citations, the world wide web or any other hypermedia database.

BACKGROUND OF THE INVENTION

Due to the developments in computer technology and its increase in popularity, large numbers of people have recently started to frequently search huge databases. For example, internet search engines are frequently used to search the entire world wide web. Currently, a popular search engine might execute over 30 million searches per day of the indexable part of the web, which has a size in excess of 500 Gigabytes. Information retrieval systems are traditionally judged by their precision and recall. What is

2

often neglected, however, is the quality of the results produced by these search engines. Large databases of documents such as the web contain many low quality documents. As a result, searches typically return hundreds of irrelevant or unwanted documents which camouflage the few relevant ones. In order to improve the selectivity of the results, common techniques allow the user to constrain the scope of the search to a specified subset of the database, or to provide additional search terms. These techniques are most effective in cases where the database is homogeneous and already classified into subsets, or in cases where the user is searching for well known and specific information. In other cases, however, these techniques are often not effective because each constraint introduced by the user increases the chances that the desired information will be inadvertently eliminated from the search results.

Search engines presently use various techniques that attempt to present more relevant documents. Typically, documents are ranked according to variations of a standard vector space model. These variations could include (a) how recently the document was updated, and/or (b) how close the search terms are to the beginning of the document. Although this strategy provides search results that are better than with no ranking at all, the results still have relatively low quality. Moreover, when searching the highly competitive web, this measure of relevancy is vulnerable to "spamming" techniques that authors can use to artificially inflate their document's relevance in order to draw attention to it or its advertisements. For this reason search results often contain commercial appeals that should not be considered a match to the query. Although search engines are designed to avoid such ruses, poorly conceived mechanisms can result in disappointing failures to retrieve desired information.

Hyperlink Search Engine, developed by IDD Information Services, (http://rankdex.gari.com/) uses backlink information (i.e., information from pages that contain links to the current page) to assist in identifying relevant web documents. Rather than using the content of a document to determine relevance, the technique uses the anchor text of links to the document to characterize the relevance of a document. The idea of associating anchor text with the page the text points to was first implemented in the World Wide Web Worm (Oliver A. McBryan, GENVL and WWWW: Tools for Taming the Web, First International Conference on the World Wide Web, CERN, Geneva, May 25-27, 1994). The Hyperlink Search Engine has applied this idea to assist in determining document relevance in a search. In particular, search query terms are compared to a collection of anchor text descriptions that point to the page, rather than to a keyword index of the page content. A rank is then assigned to a document based on the degree to which the search terms match the anchor descriptions in its backlink documents.

The well known idea of citation counting is a simple method for determining the importance of a document by counting its number of citations, or backlinks. The citation rank r(A) of a document which has n backlink pages is simply

$$r(A) = n.$$

In the case of databases whose content is of relatively uniform quality and importance it is valid to assume that a highly cited document should be of greater interest than a document with only one or two citations. Many databases, however, have extreme variations in the quality and importance of documents. In these cases, citation ranking is overly simplistic. For example, citation ranking will give the same rank to a document that is

4

cited once on an obscure page as to a similar document that is cited once on a well-known and highly respected page.

SUMMARY

~~OBJECTS AND ADVANTAGES OF THE INVENTION~~

Various aspects ~~Accordingly, it is a primary object~~ of the present invention ~~to~~ provide systems and ~~a~~ methods for ranking documents in a linked database. One aspect ~~It is another object of the invention to provide such a method that~~ provides an objective ranking based on the relationship between documents. Another aspect ~~object~~ of the invention is directed ~~to provide~~ a technique for ranking documents within a database whose content has a large variation in quality and importance. Another aspect ~~object~~ of the present invention is to provide a document ranking method that is scalable and can be applied to extremely large databases such as the world wide web. Additional aspects of the invention ~~objects and advantages~~ will become apparent in view of the following description and associated figures.

~~SUMMARY OF THE INVENTION~~

One aspect of the ~~The~~ present invention is directed to ~~achieves the above objects by~~ taking advantage of the linked structure of a database to assign a rank to each document in the database, where the document rank is a measure of the importance of a document. Rather than determining relevance only from the intrinsic content of a document, or from the anchor text of backlinks to the document, a ~~the present~~ method consistent with the invention determines importance from the extrinsic relationships between documents. Intuitively, a document should be important (regardless of its content) if it is highly cited by other documents. Not all citations, however, are necessarily of equal significance. A citation from an important document is more important than a citation from a relatively unimportant document. Thus, the importance of a page, and hence the rank assigned to it, should depend not just on the number of citations it has, but on the importance of the citing documents as well. This implies a recursive definition of rank: the rank

5

of a document is a function of the ranks of the documents which cite it. The ranks of documents may be calculated by an iterative procedure on a linked database.

Because citations, or links, are ways of directing attention, the important documents correspond to those documents to which the most attention is directed. Thus, a high rank indicates that a document is considered valuable by many people or by important people. Most likely, these are the pages to which someone performing a search would like to direct his or her attention. Looked at another way, the importance of a page is directly related to the steady-state probability that a random web surfer ends up at the page after following a large number of links. Because there is a larger probability that a surfer will end up at an important page than at an unimportant page, this method of ranking pages assigns higher ranks to the more important pages.

In one aspect of the invention, a computer implemented method is provided for ~~calculating an importance rank for N linked nodes of a~~ scoring documents linked ~~database~~. The method comprises the steps of:

~~(a) selecting an initial N-dimensional vector $\mathbf{p}_0$;~~

~~(b) computing an approximation $\mathbf{p}_n$ to a steady-state probability~~ $\mathbf{p}_\infty$ in accordance with the equation $\mathbf{p}_n = \mathbf{A}^n\mathbf{p}_0$, where $\mathbf{A}$ is an NxN transition probability matrix having elements $\mathbf{A}[i][j]$ representing a probability of moving from node i to node j; and

(c) determining a rank r[k] for a node k from a k^{th} component of $\mathbf{p}_n$.

In a preferred embodiment, the matrix $\mathbf{A}$ is chosen so that an importance rank of a node is calculated, in part, from a weighted sum of importance ranks of backlink nodes of the node, ~~where each of the backlink nodes is weighted in dependence upon~~

the total number of links in the backlink node. In addition, the importance rank of a node is calculated, in part, from a constant α representing the probability that a surfer will randomly jump to the node. The importance rank of a node can also be calculated, in part, from a measure of distances between the node and backlink nodes of the node. The initial N-dimensional vector $\mathbf{p}_0$ may be selected to represent a uniform probability distribution, or a non-uniform probability distribution which gives weight to a predetermined set of nodes.

BRIEF DESCRIPTION OF THE DRAWINGS

Fig. **1** is a diagram of the relationship between three linked hypertext documents according to the invention.

Fig. **2** is a diagram of a three-document web illustrating the rank associated with each document in accordance with the present invention.

Fig. 3 is a flowchart of one embodiment of the invention

DETAILED DESCRIPTION

Although the following detailed description contains many specifics for the purposes of illustration, anyone of ordinary skill in the art will appreciate that many variations and alterations to the following details are within the scope of the invention. Accordingly, the following ~~preferred~~ embodiments of the invention ~~is~~ are set forth without any loss of generality to, and without imposing limitations upon, the claimed invention. For support in reducing the present invention to practice, the inventor acknowledges Sergey Brin, Scott Hassan, Rajeev Motwani, Alan Steremberg, and Terry Winograd.

A linked database (i.e. any database of documents containing mutual citations, such as the world wide web or other hypermedia archive, a dictionary or thesaurus, and a database of academic articles, patents, or court cases) can be represented as a directed graph of N nodes, where each node corresponds to a web

page document and where the directed connections between nodes correspond to links from one document to another. A given node has a set of forward links that connect it to children nodes, and a set of backward links that connect it to parent nodes. FIG. 1 shows a typical relationship between three hypertext documents A, B, and C. As shown in this particular figure, the first links in documents B and C are pointers to document A. In this case we say that B and C are backlinks of A, and that A is a forward link of B and of C. Documents B and C also have other forward links to documents that are not shown.

Although the ranking method of the present invention is superficially similar to the well known idea of citation counting, the present method is more subtle and complex than citation counting and gives far superior results. In a simple citation ranking, the rank of a document A which has n backlink pages is simply

$$r(A) = n.$$

According to one embodiment of the present method of ranking, the backlinks from different pages are weighted differently and the number of links on each page is normalized. More precisely, the rank of a page A is defined according to the present invention as

T.0080

$$r(A) = \frac{\alpha}{N} + (1-\alpha)\left(\frac{r(B_1)}{|B_1|} + \ldots + \frac{r(B_n)}{|B_n|}\right),$$

where $B_1, \ldots, B_n$ are the backlink pages of A, $r(B_1), \ldots, r(B_n)$ are their ranks, $|B_1|, \ldots, |B_n|$ are their numbers of forward links, and α is a constant in the interval [0,1], and N is the total number of pages in the web. This definition is clearly

8

more complicated and subtle than the simple citation rank. Like the citation rank, this definition yields a page rank that increases as the number of backlinks increases. But the present method considers a citation from a highly ranked backlink as more important than a citation from a lowly ranked backlink (provided both citations come from backlink documents that have an equal number of forward links). In the present invention, it is possible, therefore, for a document with only one backlink (from a very highly ranked page) to have a higher rank than another document with many backlinks (from very low ranked pages). This is not the case with simple citation ranking.

The ranks form a probability distribution over web pages, so that the sum of ranks over all web pages is unity. The rank of a page can be interpreted as the probability that a surfer will be at the page after following a large number of forward links. The constant α in the formula is interpreted as the probability that the web surfer will jump randomly to any web page instead of following a forward link. The page ranks for all the pages can be calculated using a simple iterative algorithm, and corresponds to the principal eigenvector of the normalized link matrix of the web, as will be discussed in more detail below.

In order to illustrate the present method of ranking, consider the simple web of three documents shown in FIG. 2. For simplicity of illustration, we assume in this example that r=0. Document A has a single backlink to document C, and this is the only forward link of document C, so

$$r(A) = r(C).$$

Document B has a single backlink to document A, but this is one of two forward links of document A, so

$r(B) = r(A)/2$.

Document C has two backlinks. One backlink is to document B, and this is the only forward link of document B. The other backlink is to document A via the other of the two forward links from A. Thus

$r(C) = r(B) + r(A)/2$.

In this simple illustrative case we can see by inspection that $r(A) = 0.4$, $r(B) = 0.2$, and $r(C) = 0.4$. Although a typical value for α is ~0.1, if for simplicity we set $\alpha = 0.5$ (which corresponds to a 50% chance that a surfer will randomly jump to one of the three pages rather than following a forward link), then the mathematical relationships between the ranks become more complicated. In particular, we then have

$r(A) = 1/6 + r(C)/2$,
$r(B) = 1/6 + r(A)/4$, and
$r(C) = 1/6 + r(A)/4 + r(B)/2$.

The solution in this case is $r(A) = 14/39$, $r(B) = 10/39$, and $r(C) = 15/39$.

In practice, there are millions of documents and it is not possible to find the solution to a million equations by inspection. Accordingly, in the preferred embodiment a simple iterative procedure is used. As the initial state we may simply set all the ranks equal to 1/N. The formulas are then used to calculate a new set of ranks based on the existing ranks. In the case of millions of documents, sufficient convergence typically takes on the order of 100 iterations. It is not always necessary or even desirable, however, to calculate the rank of every page with high precision. Even approximate rank

values, using two or more iterations, can provide very valuable, or even superior, information.

The iteration process can be understood as a steady-state probability distribution calculated from a model of a random surfer. This model is mathematically equivalent to the explanation described above, but provides a more direct and concise characterization of the procedure. The model includes (a) an initial N-dimensional probability distribution vector $\mathbf{p}_0$ where each component $\mathbf{p}_0[i]$ gives the initial probability that a random surfer will start at a node i, and (b) an NxN transition probability matrix **A** where each component **A**[i][j] gives the probability that the surfer will move from node i to node j. The probability distribution of the graph after the surfer follows one link is $\mathbf{p}_1 = \mathbf{A}\mathbf{p}_0$, and after two links the probability distribution is $\mathbf{p}_2 = \mathbf{A}\mathbf{p}_1 = \mathbf{A}^2\mathbf{p}_0$. Assuming this iteration converges, it will converge to a steady-state probability

$$\mathbf{p}_\infty = \lim_{n\to\infty} \mathbf{A}^n\mathbf{p}_0,$$

which is a dominant eigenvector of **A**. The iteration circulates the probability through the linked nodes like energy flows through a circuit and accumulates in important places. Because pages with no links occur in significant numbers and bleed off energy, they cause some complication with computing the ranking. This complication is caused by the fact they can add huge amounts to the "random jump" factor. This, in turn, causes loops in the graph to be highly emphasized which is not generally a desirable property of the model. In order to address this problem, these childless pages can simply be removed from the model during the iterative stages, and added back in after the iteration is complete. After the childless

pages are added back in, however, the same number of iterations that was required to remove them should be done to make sure they all receive a value. (Note that in order to ensure convergence, the norm of $\mathbf{p}_i$ must be made equal to 1 after each iteration.) An alternate method to control the contribution of the childless nodes is to only estimate the steady state by iterating a small number of times.

The rank r[i] of a node i can then be defined as a function of this steady-state probability distribution. For example, the rank can be defined simply by $r[i] = \mathbf{p}_\infty[i]$. This method of calculating rank is mathematically equivalent to the iterative method described first. Those skilled in the art will appreciate that this same method can be characterized in various different ways that are mathematically equivalent. Such characterizations are obviously within the scope of the present invention. Because the rank of various different documents can vary by orders of magnitude, it is convenient to define a logarithmic rank

$$r[i] = \log \frac{\mathbf{p}_\infty[i]}{\min_{k \in [1,N]} \{\mathbf{p}_\infty[k]\}}$$

which assigns a rank of 0 to the lowest ranked node and increases by 1 for each order of magnitude in importance higher than the lowest ranked node.

In ~~a preferred~~ one particular embodiment, a finite number of iterations are performed to approximate p_∞. The initial distribution can be selected to be uniform or non-uniform. A uniform distribution would set each component of p_0 equal to 1/N. A non-uniform distribution, for example, can divide the initial probability among a few nodes which are known a priori to have relatively

large importance. This non-uniform distribution decreases the number of iterations required to obtain a close approximation to p_∞ and also is one way to reduce the effect of artificially inflating relevance by adding unrelated terms.

E In another particular ~~a preferred~~ embodiment, the transition matrix **A** is given by

T, 0130 $$\mathbf{A} = \frac{\alpha}{N}\,\mathbb{1} + (1-\alpha)\mathbf{B},$$

where $\mathbb{1}$ is an NxN matrix consisting of all 1s, α is the probability that a surfer will jump randomly to any one of the N nodes, and **B** is a matrix whose elements **B**[i][j] are given by

T, 0131 $$\mathbf{B}[i][j] = \begin{cases} \frac{1}{n_i} & \text{if node i points to node j} \\ 0 & \text{otherwise} \end{cases},$$

where n_i is the total number of forward links from node i. The $(1-\alpha)$ factor acts as a damping factor that limits the extent to which a document's rank can be inherited by children documents. This models the fact that users typically jump to a different place in the web after following a few links. The value of α is typically around 15%. Including this damping is important when many iterations are used to calculate the rank so that there is no artificial concentration of rank importance within loops of the web. Alternatively, one may set α=0 and only iterate a few times in the calculation.

E Consistent with the present invention, there ~~There~~ are several ways that this method can be adapted or altered for various purposes. As already mentioned above, rather than including the random linking probability α equally

among all nodes, it can be divided in various ways among all the sites by changing the $\mathbb{1}$ matrix to another matrix. For example, it could be distributed so that a random jump takes the surfer to one of a few nodes that have a high importance, and will not take the surfer to any of the other nodes. This can be very effective in preventing deceptively tagged documents from receiving artificially inflated relevance. Alternatively, the random linking probability could be distributed so that random jumps do not happen from high importance nodes, and only happen from other nodes. This distribution would model a surfer who is more likely to make random jumps from unimportant sites and follow forward links from important sites. A modification to avoid drawing unwarranted attention to pages with artificially inflated relevance is to ignore local links between documents and only consider links between separate domains. Because the links from other sites to the document are not directly under the control of a typical web site designer, it is then difficult for the designer to artificially inflate the ranking. A simpler approach is to weight links from pages contained on the same web server less than links from other servers. Also, in addition to servers, internet domains and any general measure of the distance between links could be used to determine such a weighting.

Additional modifications can further improve the performance of this method.Rank can be increased for documents whose backlinks are maintained by different institutions and authors in various geographic locations. Or it can be increased if links come from unusually important web locations such as the root page of a domain.

Links can also be weighted by their relative importance within a document. For example, highly visible links that are near the top of a document can be given more weight. Also, links that are

in large fonts or emphasized in other ways can be given more weight. In this way, the model better approximates human usage and authors' intentions. In many cases it is appropriate to assign higher value to links coming from pages that have been modified recently since such information is less likely to be obsolete.

Sub E2

~~The present method has~~ the advantage that the convergence is very fast (a few hours using current processors) and it is much less expensive than building a full-text index. This speed allows the ranking to be customized or personalized for specific users. For example, a user's home page and/or bookmarks can be given a large initial importance, and/or a high probability of a random jump returning to it. This high rating essentially indicates to the system that the person's homepage and/or bookmarks does indeed contain subjects of importance that should be highly ranked. This procedure essentially trains the system to recognize pages related to the person's interests.

The present method of determining the rank of a document can also be used to enhance the display of documents. In particular, each link in a document can be annotated with an icon, text, or other indicator of the rank of the document that each link points to. Anyone viewing the document can then easily see the relative importance of various links in the document.

The present method of ranking documents in a database can also be useful for estimating the amount of attention any document receives on the web since it models human behavior when surfing the web. Estimating the importance of each backlink to a page can be useful for many purposes including site design, business arrangements with the backlinkers, and marketing. The effect of potential changes to the hypertext structure can be evaluated by adding them to the link structure and recomputing the ranking.

Real usage data, when available, can be used as a starting point for the model and as the distribution for the alpha factor. This can allow this ranking model to fill holes in the usage data, and provide a more accurate or comprehensive picture.. Thus, although this method of ranking does not necessarily match the actual traffic, it nevertheless measures the degree of exposure a document has throughout the web.

Another ~~Perhaps the most~~ important application and embodiment of the present ~~ranking technique is to enhance~~ invention is directed to enhancing the quality of results from web search engines. In this application of the present invention, ~~the~~ a ranking method ~~of~~ according to the invention is integrated into a web search engine to produce results far superior to existing methods in quality and performance. A search engine employing ~~the~~ a ranking method of the present invention ~~has all the advantages of~~ provides automation while producing results comparable to a human maintained categorized system. In this approach, a web crawler explores the web and creates an index of the web content, as well as a directed graph of nodes corresponding to the structure of hyperlinks. The nodes of the graph (i.e. pages of the web) are then ranked according to importance ~~according to the method~~ of the present invention.

The search engine is used to locate documents that match the specified search criteria, either by searching full text, or by searching titles only. In addition, the search can include the anchor text associated with backlinks to the page. This ~~idea~~ approach has several advantages in this context. First, anchors often provide more accurate descriptions of web pages than the pages themselves. Second, anchors may exist for images, programs, and other objects that cannot be indexed by a text-based search engine. This also makes it possible to return web pages which have not actually been crawled. In addition, the engine can

compare the search terms with a list of its backlink document titles. Thus, even though the text of the document itself may not match the search terms, if the document is cited by documents whose titles or backlink anchor text match the search terms, the document will be considered a match. In addition to or instead of the anchor text, the text in the immediate vicinity of the backlink anchor text can also be compared to the search terms in order to improve the search.

Once a set of documents is identified that match the search terms, the list of documents is then sorted with high ranking documents first and low ranking documents last. The ranking in this case is ~~defined as~~ a function which combines all of the above factors such as the objective ranking and textual matching. If desired, the results can be grouped by category or site as well.

It will be clear to one skilled in the art that the above embodiments may be altered in many ways without departing from the scope of the invention. Accordingly, the scope of the invention should be determined by the following claims and their legal equivalents.